Alignment of Monolingual Corpus by Reduction of the Search Space (old version)

Prajol Shrestha

Alignment of Monolingual Corpus by Reduction of the Search Space

Prajol Shrestha
Prajol.Shrestha@etu.univ-nantes.fr

Résumé. Les corpus comparables monolingues, alignés non pas au niveau des documents mais au niveau d'unités textuelles plus fines (paragraphe, phrases, etc.), sont utilisés dans diverses applications de traitement automatique des langues comme par exemple en détection de plagiat. Mais ces types de corpus ne sont pratiquement pas disponibles et les chercheurs sont donc obligés de les construire et de les annoter manuellement, ce qui est un travail très fastidieux et coûteux en temps. Dans cet article, nous présentons une méthode, composée de deux étapes, qui permet de réduire ce travail d'annotation de segments de texte. Cette méthode est évaluée lors de l'alignement de paragraphes provenant de dépêches en langue anglaise issues de diverses sources. Les résultats obtenus montrent un apport considérable de la méthode en terme de réduction de temps d'annotation. Nous présentons aussi des premiers résultats obtenus à l'aide de simples traitements automatiques (recouvrement de mots, de racines, mesure cosinus) pour tenter de diminuer encore la charge de travail humaine.

Abstract. Monolingual comparable corpora annotated with alignments between text segments (paragraphs, sentences, etc.) based on similarity are used in a wide range of natural language processing applications like plagiarism detection, information retrieval, summarization and so on. The drawback wanting to use them is that there aren't many standard corpora which are aligned. Due to this drawback, the corpus is manually created, which is a time consuming and costly task. In this paper, we propose a method to significantly reduce the search space for manual alignment of the monolingual comparable corpus which in turn makes the alignment process faster and easier. This method can be used in making alignments on different levels of text segments. Using this method we create our own gold corpus aligned on the level of paragraph, which will be used for testing and building our algorithms for automatic alignment. We also present some experiments for the reduction of search space on the basis of stem overlap, word overlap, and cosine similarity measure which help us automatize the process to some extent and reduce human effort for alignment.

Mots-clés : corpus comparable monolingue, alignement, similarité.

Keywords: monolingual comparable corpus, alignment, similarity.

1 Introduction

Monolingual comparable corpora from its name can be understood to be a collection of electronic text documents in a single language collected on the basis of comparability. The characteristics of comparability has not been explained and analyzed in many literature and the ones that explain are made for bilingual comparable corpora (Maia, 2003). Comparability depends on the application for which the documents are collected for instance, for information retrieval, comparable corpora would be documents that are related to a set of other documents whereas for detecting and measuring reused text, the comparable corpora are documents that have been rewritten from a set of previously written documents (Gaizauskas *et al.*, 2001).

In the field of Natural Language Processing (NLP), monolingual comparable corpora are used to build and test a wide range of applications such as information retrieval, summarization, plagiarism detection, dictionary building and so on. With a number of application to be built, the field of NLP requires a wide range of monolingual comparable corpus with different annotations. The annotations can be at different levels starting from word to document level and different types of annotations such as Part-of-Speech annotation of words to annotations of related documents. In this paper, our focus of annotation is the annotation of alignment between text segments based on similarity and from here onwards annotation will refer to alignment. There are some standard monolingual comparable corpus available, created manually by collecting existing documents written by humans or artificially by generating machine made text. These corpora are annotated on the level of :

– Documents, TDT corpus [1], for Topic detection and tracking applications created manually.
– Text segments, PAN-PC-09 (Barrón-Cedeño *et al.*, 2010), for plagiarism detection and are artificially created due to the fact that natural text on plagiarism are hard to collect.
– Documents as well as text segments, METER corpus (Gaizauskas *et al.*, 2001), for the detection of text reuse created manually.

From the list above, we can realize that for each NLP task, that use comparable corpus, we will require a corpus with specific annotations. There aren't many corpus like these that are available and as the field of NLP expands, more corpus with specific annotations would be required. Other NLP activities that do not make use of these annotations made in the standard monolingual comparable corpus make their own specific annotations. These corpus are not available for the public and therefore for NLP tasks that does not have a standard annotated corpus, a new corpus with annotations has to be built.

Most of the annotation are made manually by two or more professional annotators who search for alignments on all possible alignment pairs. In monolingual corpora, the alignments annotated are very few compared to the total possible alignment searched by the annotators. We propose a method to reduce this search space of the total possible alignments so that this task becomes much less time consuming and costly. This method can be used to align a wide range of text segments as we show how it helps in aligning similar paragraphs in monolingual comparable corpus and also present some methods that will help automatize it to some extent.

We will start in section 2 by describing the alignment process in which the reduction of space is explained in section 2.1 and the creation of the gold corpus in section 2.2. In section 3, we present the experiments carried out and their results. We finally conclude and present our future work in section 4.

2 Alignment

Alignment, in general, could be seen as a problem in which segments of text are grouped or linked to each other based on some relations like synonymy, translation, similarity and so on (Indurkhya & Damerau, 2010). The granularity of the segments for alignment may vary : characters, words, phrases, sentences, paragraphs, or documents. The relation for alignment and the granularity of the segments depend on the application for which alignment is being done. Alignment is used in many different applications of NLP : machine translation, dictionary building, summarization, information retrieval and many more.

Our alignment process focuses between text segments within a document, for instance paragraphs or sentences, and is based on the similarity between these text segments. The alignment of these text segments is a tedious work even for a corpus containing few hundred of text segments. For *n* number of text segments the total number of comparisons between them, to decide if an alignment exists or not, equals to :

$$\frac{n(n-1)}{2} \tag{1}$$

For instance, the corpus we use has 239 paragraphs, explained in section 3.1, for alignment which makes a total of 28,441 compar-

1. http ://projects.ldc.upenn.edu/TDT-Pilot/

isons. Usually, all of these comparisons are done manually by two or more annotators which is time consuming and therefore, we propose a method to reduce this effort by reducing the number of comparisons before giving them to the annotators.

2.1 Reducing the search space

Annotating is a tedious task searching for alignments through all the possible pairs of text unit. To reduce the amount of search we make the alignment process in two phases by breaking the problem into two parts. The first phase is the part in which the total combination of alignment pairs are reduced by selecting candidate alignment pairs such that the actual alignment pairs are the subset of this candidate alignment pairs.

$$ActualAlignmentPairs \subset CandidateAlignmentPairs \tag{2}$$

With these candidate alignments the annotators have a smaller set of paragraph pairs to work with and can be more effective as well as efficient. In the next phase, the annotators select the actual alignments from these candidate alignment pairs. This is possible because in comparable corpora there are many text segment pairs that are not similar and therefore, it is possible to filter these pairs of text segments.

Selecting the candidate alignment pairs is the first phase where, we select all the pairs of text segments that has the possibility of being similar and therefore aligned. The main objective of this phase is to reduce the size of the initial total amount of alignment pairs in such a way that the actual alignments are not missed out. This selection is done when a criteria is met. To follow the criteria, we first divide each text segments into entities. According to these entities the candidate alignment pairs are selected. The entities are listed and explained below :

– *Noun Entities* : These are parts of the text segments that represents the important nouns or noun phrases of the text segments. Importance depends on how much meaning does this entity bear to convey the meaning of the whole text segment.
– *Verb Entities* : These are the parts of the sentences that represents the main intransitive verbs of the text segments. Intransitive verbs are verbs that shows action of some sort but does not have a direct object (Loberger & Shoup, 2009). As the verb is related to one noun or noun phrase the importance of this verb tends to be higher for selecting similar text segments which will be evident when alignment is done.

Here is an example of two text segments that are compared using their entities :

Text Segment I :

William and Harry, with their father Prince Charles and their grandmother Queen Elizabeth, are thought likely to remain in seclusion at Balmoral Castle in Scotland until Saturday's ceremony.

Entities :
– William
– Harry
– Father Prince Charles
– grandmother Queen Elizabeth
– seclusion
– Balmoral Castle
– Scotland
– Saturday's ceremony

Similarly, we extract entities from the second text segment with which we want to compare the previous text segment. The second text segment is given below :

Text Segment II :

One is the state funeral, normally staged only for sovereigns, although the reigning king or queen, with the approval of Parliament, can order one for others. Churchill, Britain's prime minister during World War II, is one who received such treatment in 1965.

Entities :
– state funeral

– sovereigns
– reigning king or queen
– approval of Parliament
– Churchill
– Britain's prime minister
– World War II

Once this is done for both the text segments we select this alignment as one of the candidate alignments if the following condition is satisfied :

The concept of at least one entity should be common to the entity set of the text segments.

Comparing text segments I and II, we can see that we have a common element between these text segments which are the entities 'Queen Elizabeth' and 'king or queen'. These two elements are same obviously not because of common terms but because of the concept of 'queen' which is Queen Elizabeth. This concept can be easily known using the context in the paragraph text segments. This comparison of common entities in the text segments are easier to determine than to decide if these two text segments are similar or not and therefore can be done faster than the traditional method of directly finding similar pair of text segments.

As we can see that the text segments aren't similar with any logical definition of similarity, given that the text segments convey different information, yet these are selected because of our selection criteria. This criteria that we present will theoretically guarantee that the actual aligned text segment pairs will be present in the candidate alignments.

To make it clearer about the concept of the element here are some examples of concept of entities :

word	possible concepts
crashed	rammed into a wall, fatal impact
prince Charles	heir to the British throne
grief	sadness, mourn
high speed	121 mph, flying by

This method of selection is easy and could be done by a non professional annotator on different lengths of text segments. Once these candidate alignments are collected, they can be given to the annotators for annotating the alignments by finding the actual alignments. The number of candidate alignments will be less than the original combination of pair of text segments and therefore many annotators can work efficiently on the small set in less human hours. The next section describes how we created our gold corpus from the set of candidate alignments selected by our selection criteria.

2.2 Creation of Gold Corpus

We built the gold corpus annotated with aligned paragraphs to build and test algorithms for automatic alignment of paragraphs. To build the gold corpus, two annotators selected the actual alignment pairs from the candidate alignments which were selected as explained in the previous section 2.1. Two paragraphs are annotated as aligned in the gold corpus on the basis of similarity. To show how the annotators annotated the gold corpus we first define the term similarity.

Similarity is a difficult concept to define in general because the definition of this term depends on the application for which this measure is intended. Even with this difficulty, there are many similarity measures (Barron-Cedeno *et al.*, 2009) like cosine similarity measure, with which similar texts are measured but these measures do not define similarity, they rather assign a value of similarity. Here are some of the definitions of similarity between two texts :

1. Two sentences are similar if they contain at least one clause that expresses the same information. (Barzilay & Elhadad, 2003)
2. Two paragraphs are similar if they contain "common information". This was defined to be the case if the paragraphs referred to the same object and the object either (a) performed the same action in both paragraphs, or (b) was described in the same way in both paragraphs. (Hatzivassiloglou & Klavans, 2001) (Hatzivassiloglou *et al.*, 1999)
3. Two text are similar on the basis of these intuitions :(Lin, 1998)
   – **Intuition 1 :** The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.

- **Intuition 2 :** The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.
- **Intuition 3 :** The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

All the definitions presented above focus on what is common between the text segments to call them similar. This focus on what is common is also the difference between them. Definition 1 and 2 states what should be common where as definition 3 gives no information about it and therefore, is the most general definition among them. The more general the definition is, the more difficult the annotation process becomes because of the different interpretation of the definition and in turn more disagreements between annotators. Definition 1 and 2 is difficult to apply to all the paragraphs that we see in our corpus. Definition 1 is specific to sentences and paragraphs with more than one sentence cannot be considered similar on the basis of the same information within clauses. Definition 2 considers similarity on the basis of objects and there may exist paragraphs for which the information about the objects alone do not represent the meaning of the paragraph as in the following paragraph :

French television said Diana was being pursued by paparazzi when the crash occurred, and French Interior Minister Jean-Pierre Chevenement said police were questioning seven photographers as part of a criminal investigation into the accident.

In this paragraph, the object, paparazzi, doesn't perform any action nor does the description of the paparazzi as photographers represents the paragraph.

We will define our similarity definition with the intuition 1 of definition 3 by defining the term commonality. We define the term 'commonality' in the definition on the basis of common sub-topic. The explication of sub-topic is also intuitive as we define sub-topic as the main ideas that the paragraph gives. Intuition 2 considers the differences being a basis of similarity but we only require how similar they are and the differences between paragraphs could be considered as how much the paragraphs are not similar and so we ignore this intuition. Intuition 3 is partially correct as identical paragraphs are definitely similar to it's maximum as they will share the same sub-topic but we ignore it because it is possible that two non-identical paragraphs may consist of the same sub-topic. A paragraph may have more than one sub-topic and for us a minimal of one common sub-topic would make the paragraph similar. Here is an example where we compare two paragraphs for similarity :

Paragraph I (PI) :

At Kensington Palace, the flowers covered an area estimated at 50 by 30 feet. There were more flowers as well at Harrods department store, which is owned by Dodi Fayed's father, billionaire Egyptian businessman Mohamed Fayed.

Paragraph II (PII) :

Mounds of flowers marked the sidewalk near one gate at Kensington Palace, where Diana resided, and along the main gates of the palace. There were flowers as well at Buckingham Palace, at St. James' Palace, at Harrod's department store, which is owned by Fayed's father, the Egyptian born business tycoon Mohamed Al Fayed. There were even flowers outside the gym where Diana regularly worked out.

The main ideas from Paragraph I are :

1. Flowers were present at Kensignton Palace
2. Flowers were present at Harrods department store

The main ideas from Paragraph II are :

1. Flowers were present at Kensington Palace
2. Flowers were placed at the palace
3. Flowers were placed at Buckingham Palace
4. Flowers were present at Harrods department store
5. Flowers were present at the gym

Once we have the main ideas that are present, we try to find at least one overlap between them. In paragraphs I and II we find the following overlaps between them :

PI.1 with PII.1
PI.2 with PII.4

Here is another example of finding the similarity between paragraphs which is less intuitive at the first glance : Paragraphs to compare are :

Paragraph III (PIII) :

Dodi Al Fayed's father, Harrods Department Store owner Mohammed Al Fayed, arrived here immediately after learning of his son's death.

Paragraph IV (PIV) :

Bernard Dartevelle, a lawyer for Mohamed Al Fayed, Dodi Fayed's wealthy businessman father and also the owner of the Hotel Ritz, said the revelation "changes absolutely nothing." He spoke of an "ambience of harassment" created around Diana and Fayed by the constant presence of paparazzi.

The main ideas from Paragraph III are :

– Dodi Al Fayed's father arrived here after learning his son's death.

The main ideas from Paragraph IV are :

– Bernard Dartevelle said the revelation "changes absolutely nothing."
– He spoke of an "ambience of harassment" created around Diana and Fayed by the constant presence of paparazzi.

In these two paragraphs, there is no common idea and therefore they are not selected for the actual alignment. In all of the four paragraphs in the examples given above, there is an information that are in common about Dodi Al Fayed's father but is not placed as the main idea because we believe these informations are present to support the main idea given by the paragraph and not the main idea itself. For this paper, we consider our similarity to be a binary relationship showing two paragraphs have at least one sub-topic in common or none. This binary measure can be easily changed into a continuous measure as the number of sub-topic present in the paragraphs can be counted.

3 Experiments and Results

3.1 Corpus

The corpus we used to run our experiments are taken from the Linguistic Data Consortium, LDC [2]. LDC is an organization that has a collection of a wide range of corpus for different purposes. We have selected the LDC's North American News Text Corpus [3] which is a monolingual corpus that consists of news articles from a spectrum of sources which includes New York Times ,Word Press, Associated Press, Washington post and some more. Among all these articles we selected 12 articles which were published within two consecutive days and which share the same topic to make a small monolingual comparable corpus. Our characteristics of comparability lies on the topic of the articles selected. The topic is the death of Princess Diana. These articles are from The Washington Post, Los Angeles Times, and New York Times. In these 12 articles, there are in total 239 paragraphs which will be aligned to each other on the basis of similarity.

3.2 Manual Alignment

We have manually aligned 28,441 paragraph pairs that have been selected from the corpus as explained in the previous section. The alignment process is of two parts, the first in which we select the candidate alignments and in the second part we select the actual

2. http ://www.ldc.upenn.edu/Catalog/byType.jsp
3. LDC Catalog number : LDC95T21

alignments from the candidate alignments. The first part of the alignment process was done by a single annotator to reduce the total inital alignments to only 3,416 candidate alignments. In this phase, the annotator can be flexible to decide if a candidate alignment pair is really helpful to be a candidate alignment pair or not. If this decision can be taken easily and with out doubt then the candidate alignments that is valid by our selection criteria can be ignored. This flexibility is possible because of the fact that some paragraphs may have some common concepts of the entities and yet not have anything in common other than that and as it is manually done the annotator can decide not to align them. This phase is easy and therefore fast to do. It took about 71 hours to find the set of 3,416 candidate alignments from the set of 28,441 paragraph pairs.

The second phase of finding the actual alignment was done by two annotators independently and any differences among these selection of alignments were discussed together and a decision was taken with reasoning. The alignment task took about 20 hours for both annotators and a total of 429 actual alignments were selected. The total time that took us to annotate our corpus was 91 hours. If we had directly tried to find the actual alignments without phase one, with an assumption that the time taken per paragraph pair (about 21 sec) is the same as in this second phase, it would take about 166 hours. The total time saved is 75 hours of work.

These actual alignments will be used as our gold corpus for building our algorithm for similarity. We used kappa statistics (Cohen, 1960) (Carletta, 1996) to evaluate our second phase annotation. Kappa statistics is defined as :

$$k = \frac{P_A - P_E}{1 - P_E} \tag{3}$$

where P_A is the probability of two annotators agreeing in practice and P_E is the expected probability of the two annotators agreeing. In our case P_A = 0.959, P_E = 0.780 and K = 0.813, indicating the agreement on annotations are significant.

3.3 Automatizing the Alignment

Our manual alignment method is still time consuming and difficult as manual effort has to be done so we tried to use some simple automatic methods to see how they do against the manual process. We tried stem overlap, word overlap and cosine similarity measures (Barron-Cedeno *et al.*, 2009) to find the actual alignments between similar paragraphs on the corpus we manually annotated. Before the experiments on each method we removed the stop words using a stop word list and except when using word overlap method we stemmed the remaining words using a snowball stemmer [4]. We used two types of weights for the cosine similarity measure, one the frequency of the stem within the paragraph, TF, and the second one was the TF-IDF, which is used in information retrieval (Salton & McGill, 1983). We considered a pair of paragraphs to be aligned if the threshold value was crossed. Table 1 gives the best results from these methods along with their threshold value.

Methods	Threshold	Aligned from the method	Actual Alignments included
Stem Overlap	> 0	15,276	**420**
Word Overlap	> 0	12,116	407
Cosine Similarity (weight as TF)	>0.025	14,989	**420**
Cosine Similarity (weight as TF*IDF)	>0.025	7,351	376

TABLE 1 – The table gives the number of actual alignments included in the candidate alignment which were selected by the different methods along with their threshold

These methods that have been used isn't enough to automatically select the actual alignments as the precision of these methods are very low with none of them reaching a recall of 1. The best result in terms of including the actual alignments was given by the stem overlap and cosine similarity measure, which uses TF as weights, with 420 of the actual alignments retrieved. The cosine similarity measure is based on stem overlap as shown in the equation 4

$$cos(p_1, p_2) = \frac{\Sigma_{s \in p_1 \cap p_2}(TF_{s,p_1}.TF_{s,p_2})}{\sqrt{\Sigma_{s \in p_1}(TF_{s,p_1})^2.\Sigma_{s \in p_2}(TF_{s,p_2})^2}} \tag{4}$$

where, s is the stem, p is the paragraph and TF is the stem frequency of that stem. As this method is based on stem overlap it won't do better than stem overlap in terms of the number of actual alignments included but is better in improving the precision.

4. http ://snowball.tartarus.org/

Even though these methods can't be used to find actual alignments we can use them to reduce the total initial paragraph pairs that have to be checked for candidate alignments. From the table 4 we can determine that the method which uses stem overlap and cosine similarity measure, which uses TF as weights, include 420 actual alignments and can be used to reduce the search space for finding candidate alignments if the few alignments that was not included can be ignored. Considering the time taken for the first phase, if we could ignore these undetected alignments we could save a considerable amount of time as the initial set of paragraphs are almost halved and so is the time taken to select the candidate alignments.

Some of the concepts of the alignments that were not captured by the one stem overlap method are presented below in table 2 and can be seen that some similar concepts like flower and bouquets could be shown they are similar using some knowledge source such as a dictionary while others would be a complex task.

Concept I	Concept II
100 miles per hour	flying by
Flower	Bouquets
spun into the wall	a tragic end
spun into the wall	crashed
following	pursuit
William,Charles, and Queen Elizabeth	Royal Family
causal	cause

TABLE 2 – Concepts that could not be found similar using stem overlap

Looking into the actual alignments that were captured by the method of stem overlap and the cosine similarity measure, which uses TF as weights, we saw that most of the detected alignment pairs was because of the overlap of the context of the concept rather than the concept itself which puts a question about how effective this method would be while using a small size of text segments. In the news paper corpus that we use, a large portion of paragraphs have a length of a sentence which indicates that the context of the same concept within a sentence is enough for finding the candidate alignments.

4 Conclusions and Future Work

The total number of actual aligned text segments in the gold corpus shows that only 1.5% of the 28,441 initially paired paragraphs are aligned and the effort to check the other 98.5% of the paired paragraphs is wasted in terms of the difference between the end number of actual alignments and the total number of initial paired paragraphs. Our manual alignment method can reduce this wasted effort and have saved us about 75 hours of work. This method of manual alignment by reducing the search space is better than existing methods of manual annotation as the annotators have less candidate alignment pairs to annotate. These candidate alignments are easier and faster to select than the actual similar pair of text segments because of the complex nature of the definition of similarity in terms of analyzing the text according to the definition.

Even though our method is easier, it still requires much effort to select the candidate alignments. We can further reduce this effort by reducing the original set of text segment pairs using stem overlap or cosine similarity measure for choosing the candidate alignments from which the original alignments are selected to make the gold corpus. The different methods we used also showed that simple method of stem or word overlap and even cosine similarity measure are not enough to capture text segment similarities but has given a view that cosine similarity measure increases the precision. The context is also an important part in finding the actual alignments as we see with the stem overlap method that some actual alignments are captured using the context which gives us motivation in trying to use the context, like co-occurrence, in a vector space model (Kaufmann, 2000). This hypothesis will be used in the future to make our automatic alignment algorithm.

Références

BARRON-CEDENO A., EISELT A. & ROSSO P. (2009). Monolingual text similarity measures : A comparison of models over wikipedia articles revisions. In *Proceedings of the ICON : 7th International Conference on NLP*, p. 29–38.

BARRÓN-CEDEÑO A., POTTHAST M., ROSSO P., STEIN B. & EISELT A. (2010). Corpus and Evaluation Measures for Automatic Plagiarism Detection. In N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER

& D. TAPIAS, Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10)* : European Language Resources Association (ELRA).

BARZILAY R. & ELHADAD N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the conference on Empirical methods in NLP*, p. 203–212.

CARLETTA J. (1996). Assessing agreement on classification tasks : The kappa statistic. In *Computational Linguistics*, p. 249–254.

COHEN J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psyhological Measurement*, p. 37–46.

GAIZAUSKAS R., FOSTER J., WILKS Y., ARUNDEL J., CLOUGH P. & PIAO S. (2001). The meter corpus : A corpus for analysing journalistic text reuse. p. 214–223.

HATZIVASSILOGLOU V. & KLAVANS J. L. (2001). Simfinder : A flexible clustering tool for summarization. In *Proceedings of NAACL Workshop of Automati Summarization*, p. 203–212.

HATZIVASSILOGLOU V., KLAVANS J. L. & ESKIN E. (1999). Detecting text similarity over short passages : exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, p. 203–212.

INDURKHYA N. & DAMERAU F. J. (2010). *Handbook of natural language processing*. Taylor and Francis.

KAUFMANN S. (2000). Second-order cohesion. *Computational Intelligence*, **16**, 511–524.

LIN D. (1998). An information-theoretic definition of similarity. In *ICML*, p. 296–304.

LOBERGER G. & SHOUP K. (2009). *Websters New World English Grammar Handbook*. Wiley, Hoboken.

MAIA B. (2003). What are comparable corpora ? In *Proceedings of pre-conference workshop Multilingual Corpora : Linguistic Requirements and Technical perspectives, at Corpus Linguistics*, p. 27–34 : Lancaster U.K.

SALTON G. & MCGILL M. J. (1983). *Introduction to Modern Informational Retrieval*. McGraw-Hill.